Analysis of the Auditory System via Sound Texture Synthesis

McWalter, Richard Ian; Dau, Torsten

Published in:
Proceedings of the International Conference on Acoustics - AIA-DAGA 2013

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Analysis of the Auditory System via Sound Texture Synthesis

Richard McWalter, Torsten Dau[1]
Centre for Hearing and Speech Sciences, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
[1]*Email: tdau@elektro.dtu.dk*

Introduction

Humans experience many sensory inputs, from somatosensory to olfaction, allowing for the recognition of events in the surrounding environment. Fundamentally, the recognition of events is from the transduction between a sensory input and neurological signals. For sound events, the perception of auditory neurological signals allow humans to differentiate sounds where sensory inputs are processed by the brain and a specific response occurs. The peripheral processing performed by the human auditory system has incurred much interest and research; however, mid and high-level processes, which contribute towards the perception of sound, are yet to be fully understood.

Hearing research has traditionally been focused on two types of stimuli; artificial sounds and speech. Artificial sounds, such as tones and noise, have been used for hearing threshold experiments, frequency discrimination tasks, and modulation detection and have been crucial in understanding the auditory system. Speech has a significant role as well because it is the primary means of local and remote communication for many humans. The study of hearing; however, can be extended to include a broader range of stimuli. One such category is *sound textures*, which encompass a large range of naturally occurring sounds, such as birds singing or streams flowing. Sound textures are defined by their temporal homogeny and have been shown useful in the study of mid-level auditory processes [1]. The study of textures stems from visual processing research, where synthesized image textures were generated based on a set of statistics measured from a visual texture model [2, 3]. This approach of analysis via synthesis offers an alternative means of investigating sensory processing systems and uncover the critical aspects of perceptions [4].

This paper explores the way in which humans perceive natural sounds via the synthesis of textures. The investigation began with the deconstruction of the sound texture as it would occur in the peripheral auditory system. This was covered by well accepted processing stages based on both physiological and psychoacoustics data [5, 6, 7, 8]. Additional auditory stages extend the model into the mid-level processing of sound, exploring model components that contribute to the perception of natural sounds [10, 9]. The deconstructed sound sample was analyzed using a set of neurologically plausible statistics at different stages of the auditory model, capturing the differentiable features between sound textures [4]. The identified statistics were inspired from previous work in both image texture synthesis and sound texture synthesis [1, 3, 4]. The statistics gathered for a particular sound texture were then used to synthesize a comparable texture.

It is proposed that if the original natural sound texture and the synthetic sound texture possess the same statistics, they should be perceived as the same [4]. This unique approach explores the way in which the auditory system perceives natural sound textures. This method was supported with psychophysical experiments, and proved to offer compelling synthetic sound textures only when the auditory model was biologically inspired and a complete set of statistics were used. The results suggest that the perception of sound textures is based on a small set of statistics analyzed from sound processed by the auditory periphery.

Method

Sound Texture Analysis Model

The analysis of the natural sound textures was performed by a model encompassing the peripheral frequency selectivity and mid-level temporal processing characteristics of the auditory system. The first stage of the model represents the frequency selectivity of the peripheral auditory system. This is accomplished by means of 32, 4th order Gammatone filters, using equivalent rectangular band center frequency spacing from 50 Hz to 8 kHz. The second stage contains two sub-processing stages; the first sub-stage accounts for the compressive non-linearity of the basilar membrane by applying an exponential constant value of 0.3 to the output of each Gammatone filter, while the second sub-stage computes the envelope of the non-linear Gammatone filter output using a Hilbert transform. This envelope is subsequently down-sampled to a frequency of 400 Hz. The third, and final stage models the temporal processing of the auditory system. This stage processes the down-sampled non-linear envelope of each Gammatone filter output by a second filterbank, consisting of 9 filters with octave spacing. The lowest filter has a centre frequency of 1 Hz and the highest filter has a centre frequency of 200 Hz and are of filter type low-pass and high-pass, respectively. The 7 remaining filters are of type bandpass and have octave spacing between 2 Hz and 128 Hz and a constant Q-factor.

To capture the perceptually significant features of the sound textures, a set of statistics were identified at the subband envelope stage and the modulation filtering stage. The first layer of statistics includes marginal moments, comprised of the mean, variance, skewness and

kurtosis measured at the output of each subband envelope and pairwise correlation statistics between neighboring subband envelopes. The second layer of statistics is measured for the modulation filter stage, and includes modulation power measured at the output of each modulation filter and a pairwise correlation statistic measured across subband envelopes for the same modulation filter centre frequency. The analysis performed on a natural sound texture input results in 768 statistical values, regardless of the input signal duration.

Sound Texture Synthesis

A synthesis system was integrated with the analysis model in order to create new sound textures, as is shown in Figure 1. The system accepts a Gaussian noise input, which is subsequently deconstructed using a parallel analysis model, identical to that used in the sound texture analysis model. The deconstructed Gaussian noise signal is modified so that the statistics of the natural sound texture input are imposed on the noise signal. This modification of statistics is achieved by means of Gradient descent, where the error between the natural sound texture statistics and the noise input are minimized. This process is done for both the subband envelope stage and the modulation filter stage. A difference in the parallel analysis model exists in that the computation of the non-linear down-sampled subband envelope preserves an accompanying subband fine structure which is used during the reconstruction process.

Reconstruction of the modified noise input from the modulation filtering domain to the single-channel time domain signal is achieved in three stages: reconstruction from the modulation filtering stage, inversion of the non-linearity and transformation from the envelope to the up-sampled fine structure, and lastly reconstruction from the subband fine structure to a single channel time domain signal. The reconstruction of the modulation filter stage is achieved using synthesis polyphase matrices, computed from the modulation filterbank, which yields the non-linear down-sample subband envelope [11]. The non-linearity applied to the subband envelope is inverted and the resulting signal is recombined with the subband fine structure, yielding the subband envelope signals. The final reconstruction process is from the subband envelope stage to a single channel time domain signal, which is again achieved using synthesis polyphase matrices computed from the Gammatone filterbank.

In order to ensure that the statistics of the natural sound texture and synthetic sound texture are the same, a synthesis acceptance criteria was established. The modifications of statistics at both the subband envelope stage and the modulation filter stage change the bandwidth of the respective signals, which results in the reconstructed signal having slightly different statistics than the deconstructed signal. For this reason, it is necessary to impose the statistics in an iterative process. Synthesis was considered successful when the signal-to-noise ratio was greater than 20dB, where the signal is the natural sound texture statistics and the noise is the difference between the natural sound texture and synthetic sound texture statistics.

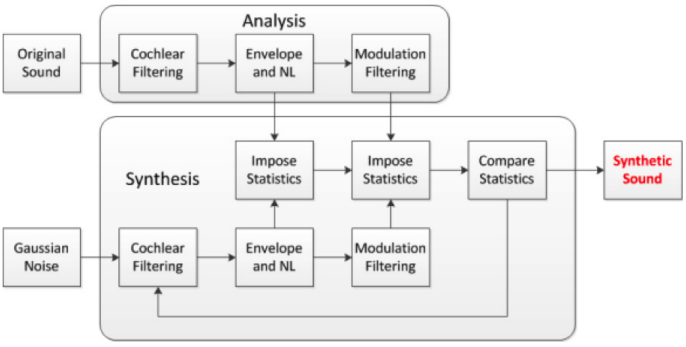

Figure 1: Sound texture synthesis overview. For details see main text.

Subjective Evaluation of Synthesized Sound Textures

Evaluation of the synthetic sound textures along with the significance of the biologically inspired auditory model and identified statistics was achieved by means of two psychophysical experiments: synthetic texture identification and modified analysis model parameters. The first experiment required the subjects to identify a sound texture from a list of 5 descriptors. This experiment was conducted using graduating statistics, from the subband envelope power to the complete statistical set. The second experiment modified analysis model parameters and examined the significance of the biologically inspired model over a modified auditory model. Four modifications were applied to the analysis model; the first was broadened peripheral Gammatone filters, the second was linearly spaced Gammatone filter centre frequencies, the third and fourth modifications dealt with to the modulation filter stage, replacing the filterbank model by a single low-pass filter with cutoff frequencies of 5 Hz or 150 Hz.

Results

Sound Texture Identification

The sound texture identification experiment results are presented in Figure 2. As more statistics were included in the synthesis of a sound texture, identification performance increased. The complete set of statistics displayed performance comparable to that of the original sound texture, indicating that the analysis system is capturing the significant features of the original sound texture.

The ability of the test subjects to correctly identify sound textures was very similar when only the peripheral subband envelope power (cochlear power) and the peripheral subband envelope marginal statistics (cochlear marginals) were included. This is demonstrated in the first two bars of Figure 2. This should not be interpreted as the marginal statistics not playing a role in characterizing sound textures. The marginal statistics capture vital information; however, in isolation, the marginal statistics push all sounds towards those

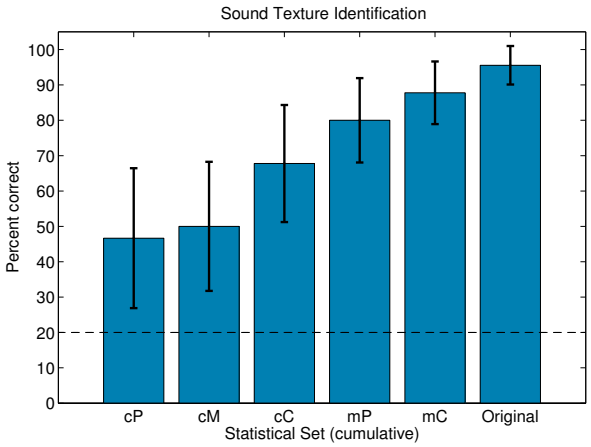

Figure 2: Identification of synthetic textures with varying statistical sets and the original textures. Note that the statistical sets are cumulative, such that the statistic descriptions along the x-axis contain all previous stages (to the left). The bars present the mean results for each statistical set, with the standard deviation presented as a black error bar. (cP = subband envelope power, cM = subband envelope marginals, cC = subband envelope correlation, mP = modulation power, mC = modulation correlation.

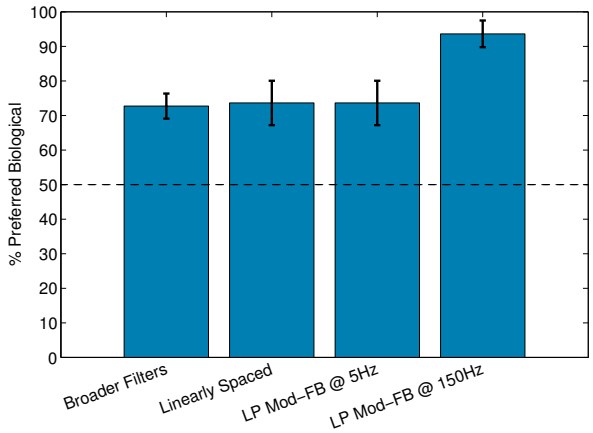

Figure 3: Percent preferred biological system results for 4 variations on auditory model. The first bars, *Broader Filters* and *Linearly Spaced*, represented deviations from the biologically inspired model at the peripheral processing stage. The third and fourth bar, *LP Mod-FB @ 5Hz* and *LP Mod-FB @ 150Hz*, represent deviations from the mid-level processing modulation filterbank processing stage.

statistics found in water sounds. The inclusion of the peripheral subband envelope correlation, which accounts for broadband events, increased the ability of the test subject to correctly identify many sound textures. The modulation power and modulation correlation statistics, accounting for temporal and modulation frequency dependent correlation features, respectively, were critical in capturing perceptually significant features found in many natural sound textures.

Modified Analysis Model Parameters

The second experiment investigated auditory models that differ from the biologically inspired model. The first part examined the peripheral filtering stage, modifying the processing with two variations of the filterbank. The results are shown in Figure 3. The first bar, which is a filterbank with broader shows that the subjects preferred sound textures synthesized with the biological model. Examining the results with more detail showed that subjects most often selected noise-like sounds, such as rain or fire, as preferred when processed using broader filters. This suggests that frequency selectivity is more significant when there is some tonal characteristic to the sound texture. The second bar presents a comparison between a biologically motivated peripheral filterbank and a linearly spaced filterbank. The results show that listeners preferred the biologically motivated peripheral filtering. Test subjects, again, appeared to prefer certain noise-like sounds with a linearly spaced filterbank.

The third and fourth bars in Figure 3 show the results from the modulation processing modification. When the modulation filterbank was exchanged with a simple low-pass filter with cut-off frequency of 5 Hz, there was a significant degradation in the synthesis of sound textures. Many textures were found to contain perceptually significant frequency dependent modulation features. These features could not be captured if modulation processing stage were modeled as a simple low-pass filter with cut-off frequency of 5 Hz, because often the modulation rates exceed this upper cutoff frequency. Increasing the cut-off frequency to 150 Hz is shown in the fourth bar, and was an attempt to preserve modulation sensitivity up to higher envelope modulation frequencies, which are present in many sound textures. This proved to be detrimental to synthesis performance as well, as the subjects preferred sound textures generated with the modulation filterbank model.

Discussion

Synthesis Statistics

The peripheral stage, comprised of the Gammatone filterbank, compressive non-linearity, and envelope extraction was captured using marginal moments and pair-wise correlation statistics. The peripheral marginal moments capture the envelope distribution of each peripheral channel. The lower marginal moments, namely the mean and the variance, reveal the power and range of the envelope, while the higher marginal moments capture the sparsity of the envelope signal. Subjectively, the marginal moments were able to capture water-like sounds, such as streams and rain. However, further work will be conducted into the significance of the peripheral marginal moments to the realism of a broader range of sound textures. The pair-wise correlation statistics capture events occurring across peripheral envelope channels. Although significant broadband events were captured with these statistics, the perceived realism of strong on-set or off-set was lacking. This suggests that additional statistics may be required to capture phase information, as have previously been identified [4].

The modulation filter was characterized, again, by

marginal statistics and pair-wise correlation. The power of each modulation band was measured and proved to be significant in capturing temporally constant events in sound textures. The modulation rates of 32 Hz and 64 Hz were particularly interesting, as they captured tonal characteristic present in the original sound texture. The pair-wise correlations, measured across subband-envelope for the same modulation filter, were also conducted. These statistics were identified as complimentary to the subband envelope correlation statistics, because they were able to highlight which modulation frequencies were correlated [4].

The means for imposing the statistics was also inspired by previous work in sound texture synthesis [4]. However, in this work, the deconstruction of the signal to the peripheral stage and modulation stage required the convergence of statistics at two levels. The gradient descent method was used at both levels, which required the analytic solution for both marginal moments and correlation statistics. The convergence of the statistics to the target SNR of 20 dB was found to occur within 30 iterations.

Auditory Processing

The auditory model used in the analysis of sound textures was guided by results from fundamental psycho-acoustics, but also modeling of more complex auditory signals [6, 9, 10]. The model uses a Gammatone filterbank, which covers the basic frequency selectivity of the peripheral auditory system. The non-linear processing of the subband signal, along with the envelope extraction, proved to be a key point of analysis for sound textures, as was shown in Figure 2. In addition, the simple statistics measured after modulation filtering introduced core temporal aspects present in many sound textures.

The modulation filterbank design follows closely that detailed in auditory models, using octave spaced filters with a constant Q-factor [9, 10]. The upper and lower frequency of the filterbank was investigated for many sound textures, and it was found that modulation frequencies of 1 Hz up to 200 Hz were perceptually significant for sound texture synthesis.

The study into non-biological auditory models, or models which do not account for modulation frequency selectivity, yield poorer synthetic textures. The results summarized in Figure 3, are significant for two reasons; firstly, broadening peripheral filters similar to those of a hearing-impaired person show degraded performance in texture synthesis and, secondly, the use of low-pass modulation filters, rather than a modulation filter bank, also degrades performance. The broadened, or hearing impaired, peripheral filterbank is valuable as it may offer a means of investigating auditory system pre-processing algorithms (i.e. assisted listening devices). The investigation into the auditory system's temporal processing demonstrated that the auditory system requires both acoustic frequency and modulation frequency sensitivity in order to differentiate sound textures.

Conclusion

The use of sound textures for exploration of the auditory system is a novel approach, which is complimentary to modern psycho-acoustic research. The identification of the key statistics that are perceptually significant provides insight into the functional analysis of sound textures by the auditory system. In addition, the basic model, which is relevant to many auditory signals, provides a strong foundation for the statistical analysis. The two psychophysical experiments support previously documented work into sound textures, but also demonstrated the significance of the modulation frequency sensitivity.

References

[1] McDermott, J.H. and Oxenham, A.J. and Simoncelli, E.P.: Sound texture synthesis via filter statistics IEEE WASPAA 2009 , 297–300

[2] Field, D.J.:Relations between the statistics of natural images and the response properties of cortical cells J. Opt. Soc. Am. A **12** (1987), 2379-2394

[3] Portilla, J. and Simoncelli, E.P.: A parametric texture model based on joint statistics of complex wavelet coefficients Int. J. Comput. Vis. **40** (2000), 49-70

[4] McDermott, Josh H and Simoncelli, Eero P.: Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis Neuron **71** (2011), 926-940

[5] Ruggero, M.A. and Rich, N.C. and Recio, A. and Narayan, S.S. and Robles, L.: Basilar-membrane responses to tones at the base of the chinchilla cochlea J. Acoust. Soc. Am. **101** (1997), 2151

[6] Patterson, RD and Nimmo-Smith, I. and Holdsworth, J. and Rice, P.: An efficient auditory filterbank based on the gammatone function APU report (1988), 2341

[7] Glasberg, B.R. and Moore, B.C.J.: Derivation of auditory filter shapes from notched-noise data Hear. Res. **47** (1990), 103-138

[8] Harte, J.M. and Elliott, S.J. and Rice, H.J.: A comparison of various nonlinear models of cochlear compression J. Acoust. Soc. Am. **117** (2005), 3777

[9] Jørgensen, S. and Dau, T.: Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing J. Acoust. Soc. Am. **130** (2011), 1475

[10] Dau, T. and Kollmeier, B. and Kohlrausch, A.: Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers J. Acoust. Soc. Am. **102** (1997), 2892

[11] Bolcskei, H. and Hlawatsch, F. and Feichtinger, H.G.: Frame-theoretic analysis of oversampled filter banks IEEE Transactions on Signal Processing **46** (1998), 3256-3268